Minimax Lower Bound: Fano's method

Tri Nguyen

nguyetr9@oregonstate.edu

August 4, 2022

Let's go with an example to have a more concrete understanding of what we are trying to analyze.

Example. Given a distribution family $\mathcal{N}_d = \left\{N(\theta, \sigma^2 I_{d\times d}) \mid \theta \in \mathbb{R}^d\right\}$. We wish to estimate $\theta(P), P \in \mathcal{N}_d$ in mean-squared error given n i.i.d samples drawn from P.

The question is what would be the best estimator that we can get in term of MSE

$$\mathbb{E}\left[\left\|\theta - \widehat{\theta}\right\|^2\right]$$

or

$$\mathcal{M}(\theta(\mathcal{N}_d), \|\cdot\|^2) := \inf_{\widehat{\theta}} \sup_{P\in\mathcal{P}} \mathbb{E}\left[\left\|\theta - \widehat{\theta}\right\|^2\right]$$

First attempt: Cramer-Rao lower bound . Funny, it is only applicable when $\theta \in \mathbb{R}$ — a scalar. For a multi-dimension $\theta \in \mathbb{R}^d$, we don't have such bound.

Likelihood function of the joint pdf is

$$f(\theta) = \frac{1}{\sigma^d\sqrt{2\pi}^d}\prod_{i=1}^{n}\exp\left(-\frac{1}{2\sigma^2}\|x_i - \theta\|^2\right)$$

$$\ln f(\theta) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\|x_i - \theta\|^2$$

$$\frac{\partial \ln f(\theta)}{\partial \theta} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \theta) = \frac{n}{\sigma^2}\left(\frac{1}{n}\sum_{i=1}^{n}x_i - \theta\right)$$

Hence the best unbiased estimator of θ is $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$, where it is the best in terms of MSE

$$\mathrm{Var}(\widehat{\theta}) = \mathbb{E}\left[\left\|\widehat{\theta} - \theta\right\|^2\right] = \frac{\sigma^2}{n}$$

abcd, it is not natural in case of multiple variables. And I wonder is there any other method to analyze/find optimal MSE?

Second attempt. When do we use unit ball, or grid space as the local packing?

First, consider

$$\mathcal{M}(\theta(\mathcal{N}_d), \Phi\circ\rho) := \inf_{\widehat{\theta}} \sup_{P\in\mathcal{P}} \mathbb{E}[\Phi(\rho(\theta, \widehat{\theta}))]$$

In order to get a bound of this $\Phi(\rho(\theta, \widehat{\theta}))$, we can imagine to partition the whole parameter θ space into a finite number of "cells". Then, the process of choosing $\widehat{\theta}$ can be reduced to choosing a "cell" with the trade of some "marginal" error within that "cell". That means, if a cell is chosen correctly, the largest error can be made is the "diameter" of that cell.

So that would be the idea of "transforming" an error analysis to hypothesis testing analysis. We have the following lemma.

Lemma 0.1. *Choose some distribution $P_v \in \mathcal{P}, v \in \mathcal{V}, |V| \leq \infty$ to represent $\mathcal{P}$, such that for $v_i \neq v_j$, we have $\rho(\theta_{v_i}, \theta_{v_j}) \geq 2\delta$. Define*

- *Let V be a RV with uniform distribution over $\mathcal{V}$,*
- *For an estimator $\widehat{\theta}$, let $\Psi(X_1^n) := \arg\min_{v \in \mathcal{V}} \rho(\theta_v, \widehat{\theta}(X_1^n))$,*

We have

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \inf_{\Psi} P(\Psi(X_1^n) \neq V)$$

- We want the RHS as large as possible.
- The partitioning does not need to cover the whole space of θ.

It seems that there is no relation between X_1^n and θ, or V.

Remark 0.1. The key in finding a tight bound is how to choose set $\mathcal{V}$ as well as δ.

- If δ is large, then $\Phi(\delta)$ would be large. But it also leads to $|\mathcal{V}|$ small, hence hypothesis testing error would be small.
- And the contrary when δ is too small.
- How to choose the right δ would depend on problem to problem.

Proof. Recall the definition of $M_n(\theta(\mathcal{P}), \Phi \circ \rho)$,

$$M_n(\theta(\mathcal{P}), \Phi \circ \rho) = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}[\Phi(\rho(\theta(P), \widehat{\theta}))]$$

$$\begin{aligned}
\mathbb{E}[\Phi(\rho(\theta, \widehat{\theta}))] &\geq \mathbb{E}\left[\Phi(\delta) I(\rho(\widehat{\theta}, \theta) \geq \delta)\right] \quad \text{(since } \Phi(x) \text{ is non-decreasing)} \\
&= \Phi(\delta) P(\rho(\theta, \widehat{\theta}) \geq \delta)
\end{aligned}$$

Let's choose a set of $|\mathcal{V}|$ candidates $P_v \in \mathcal{P}, v \in \mathcal{V}$ scattered enough such that for $v_i \neq v_j$, we have

$$\rho(\theta(P_{v_i}), \theta(P_{v_j})) \geq 2\delta$$

For a fixed estimator $\widehat{\theta}$,

$$\sup_{P \in \mathcal{P}} \mathbb{P}_{X_1^n}(\rho(\theta(P), \widehat{\theta}) \geq \delta) \geq \frac{1}{|\mathcal{V}|} \sum_v P(\rho(\theta(P_v), \widehat{\theta}) \geq \delta)$$

Now define

$$\Psi(X_1^n) := \arg\min_{v \in \mathcal{V}} \rho(\theta_v, \widehat{\theta}(X_1^n))$$

Event $\Psi(X_1^n) \neq v$ implies $\rho(\theta(P_v), \widehat{\theta}) \geq \delta$ (the other way around does not hold). Hence,

$$\mathbb{P}(\rho(\theta(P_v), \widehat{\theta}) \geq \delta) \geq \mathbb{P}(\Psi(X_1^n) \neq v)$$

which leads to

$$\sup_{P \in \mathcal{P}} \mathbb{P}_{X_1^n}(\rho(\theta(P), \widehat{\theta}) \geq \delta) \geq \frac{1}{|\mathcal{V}|} \sum_v \mathbb{P}(\Psi(X_1^n) \neq v) = \mathbb{P}(\Psi(X_1^n) \neq V)$$

where the last equality hold only if we assume V is a RV uniformly distributed over $\mathcal{V}$, which is in our control.

Lastly, taking infimum over Ψ, we get our conclusion

$$\begin{aligned}
\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}[\Phi(\rho(\theta, \widehat{\theta}))] &\geq \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \Phi(\delta) \mathbb{P}(\rho(\theta, \widehat{\theta}) \geq \delta) \\
&\geq \Phi(\delta) \inf_{\widehat{\theta}} \mathbb{P}(\Psi(X_1^n) \neq V) \\
&= \Phi(\delta) \inf_{\Psi} \mathbb{P}(\Psi(X_1^n) \neq V) \quad \text{(by the definition of } \Psi)
\end{aligned}$$

□

Next, we need to turn the hypothesis testing error into something that we can compute.

1 Local Fano

Lemma 1.1 (From Fano inequality)**.**

$$\inf_{\Psi} \mathbb{P}(\Psi(X_1^n) \neq V) \geq 1 - \frac{I(V; X_1^n) + \log 2}{\log |\mathcal{V}|}$$

Lemma 1.2 (Mutual Information to KL)**.**

$$\begin{aligned} I(V; X) &= D_{\text{kl}}(\mathbb{P}_{X,V} || \mathbb{P}_X \mathbb{P}_V) \\ &= \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} D_{\text{kl}}(P_v || \overline{P}) \quad \textit{(thanks to the uniform of } V\textit{)} \\ &\leq \frac{1}{|\mathcal{V}|^2} \sum_{v, v'} D_{\text{kl}}(P_v || P_{v'}) \end{aligned}$$

Remark 1.1. Combine all these lemmas, we get

$$\mathcal{M}(\theta(\mathcal{P}, \Phi \circ \rho)) \geq \Phi(\delta) \left(1 - \frac{I(V; X_1^n) + \log 2}{\log |\mathcal{V}|}\right)$$

So the trick is to choose $\mathcal{V}$ so that it is both 2δ-packing in order to apply the first lemma (transforming to hypothesis testing) and

$$D_{\text{kl}}(P_v || P_{v'}) \leq C\delta^2 \quad \text{for all } v, v' \in \mathcal{V}$$

in order to bound the RHS.

We want to find a set $\mathcal{V}$ that contains scattered elements, but the "diameter" cannot be too large, and number of elements also needs to be high enough.

1.1 Example

Example 1 (Normal mean estimation) Given the family $\mathcal{N}_d = \left\{N(\theta; \sigma^2 I_d) | \theta \in \mathbb{R}^d\right\}$. The task is to estimate the mean $\theta(P)$ for some $P \in \mathcal{N}_d$ given n i.i.d samples drawn from P. We wish to find out the minimax error for this in terms of mean-squared error.

Let's construct the "local packing" set $\mathcal{V}$:

- Let $\mathcal{V}_0$ be a 1/2-packing of the unit l_2-ball with cardinality of at least 2^d. The existence of this $\mathcal{V}$ is guaranteed by Lemma [xxx].
- Our local packing would be $\mathcal{V} = \left\{\delta v \in \mathbb{R}^d | v \in \mathcal{V}_0\right\}$.

Then we have for any $v, v' \in \mathcal{V}, v \neq v'$,

$$\|\theta_v - \theta_{v'}\| = \delta \left\|v - v'\right\| \geq \frac{\delta}{2} \quad \text{(since } \mathcal{V}_0 \text{ is 1/2-packing)}$$

and,

$$\|\theta_v - \theta_{v'}\| \leq \delta \quad \text{(since } v, v' \text{ are in } l_2\text{-ball)}$$

Then, apply Lemma above,

$$\begin{aligned} \mathcal{M}(\theta(\mathcal{N}_d), \|\cdot\|^2) &\geq \left(\frac{1}{2}\frac{\delta}{2}\right)^2 \left(1 - \frac{I(V; X_1^n) + \log 2}{\log |\mathcal{V}|}\right) \\ &= \frac{\delta^2}{16} \left(1 - \frac{I(V; X_1^n) + \log 2}{\log |\mathcal{V}|}\right) \end{aligned}$$

Then,

$$\begin{aligned} I(V; X_1^n) &\leq \frac{1}{|\mathcal{V}|^2} \sum_{v,v'} D_{\mathrm{kl}}(P_v^n || P_{v'}^n) \quad \text{(hm?)} \\ &= \frac{1}{|\mathcal{V}|^2} \sum_{v,v'} n D_{\mathrm{kl}} \left(N(\delta v, \sigma^2 I_d), N(\delta v', \sigma^2 I_d) \right) \\ &= n D_{\mathrm{kl}} \left(N(\delta v, \sigma^2 I_d), N(\delta v', \sigma^2 I_d) \right) \\ &= n \frac{\delta^2}{2\sigma^2} \left\| v - v' \right\|^2 \\ &\leq \frac{n\delta^2}{2\sigma^2} \end{aligned}$$

Let's combine these 2 inequalities above,

$$\mathcal{M}(\theta(\mathcal{N}_d), \|\cdot\|^2) \geq \frac{\delta^2}{16} \left(1 - \frac{\frac{n\delta^2}{2\sigma^2} + \log 2}{d \log 2} \right) = \frac{1}{32 d \sigma^2 \log 2} \left(\delta^2 (2\sigma^2 (d-1) \log 2 - n\delta^2) \right)$$

That bound's optimal value is achieved at $\delta^2 = \frac{(d-1)\sigma^2 \log 2}{n}$, and the optimal value is

$$\frac{(d-1)^2 \sigma^2 \log 2}{32 d n} \Rightarrow O\left(\frac{d\sigma^2}{n} \right)$$

We can check to see that the sample mean estimator will attain this risk's order.

We construct our local packing from the unit l_2-ball. Why? We know $\mathcal{V}_0$ can be exponentially large, but why choosing l_2-ball?

Another example?

2 A distance-based Fano method

Lemma 2.1. *Assume V is uniformly distributed over $\mathcal{V}$,*

$$\mathcal{M}(\Theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \left(1 - \frac{I(X; V) + \log 2}{\log |\mathcal{V}|} \right)$$

,

$$I(X; V) = \frac{1}{|\mathcal{V}|} \sum_v D_{\mathrm{kl}}(P_v || \overline{P}) \leq \frac{1}{|\mathcal{V}|^2} \sum_{v,v'} D_{\mathrm{kl}}(P_v || P_{v'})$$

Example: Given a P from $\mathcal{N}(\theta, \sigma^2 I_d)$, estimate θ.

Construct a 2δ-packing indexed by V: First, let $\mathcal{V}$ as a $1/2$-packing of unit l_2-ball with cardinality at least 2^d.